

Инженерная школа информационных технологий и робототехники(ИШИТР)
 Направление подготовки 09.03.01 Информатика и вычислительная техника
 Отделение информационных технологий(ОИТ)

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Предсказания повторных вызовов скорой помощи алгоритмами машинного обучения
УДК 004.85.021:654.153-022.336:614.88

Студент

Группа	ФИО	Подпись	Дата
8B5A	Мнацаканян Рафаэль Эдуардович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент отделения информационных технологий	Кудинов Антон Викторович	Кандидат технических наук		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент отделения социально-гуманитарных наук	Шулинина Юлия Игоревна			

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент отделения общетехнических дисциплин	Мезенцева Ирина Леонидовна			

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент отделения информационных технологий	Погребной Александр Владимирович	Кандидат технических наук		

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код результатов	Результат обучения (выпускник должен быть готов)	Требования ФГОС, критерии АИОР
P1	Применять базовые и специальные естественнонаучные и математические знания в области информатики и вычислительной техники, достаточные для комплексной инженерной деятельности.	Требования ФГОС (ОК-7, ОПК-5, ПК-3), критерий 5 АИОР (п. 1.1)
P2	Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач.	Требования ФГОС (ОК-7, ОПК2, 5, ПК-1, 3), критерий 5 АИОР (п.1.1, 1.2)
P3	Ставить и решать задачи комплексного анализа, связанные с созданием аппаратно-программных средств информационных и автоматизированных систем, с использованием базовых и специальных знаний, современных аналитических методов и моделей.	Требования ФГОС (ОК-6, ОПК-1, ПК-2, 4), критерий 5 АИОР (п. 1.2)
P4	Разрабатывать программные и аппаратные средства (системы, устройства, блоки, программы, базы данных и т. п.) в соответствии с техническим заданием и с использованием средств автоматизации проектирования.	Требования ФГОС (ОК-7, ОПК2, 4, ПК-1, 2), критерий 5 АИОР (п. 1.3)
P5	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания аппаратных и программных средств информационных и автоматизированных систем.	Требования ФГОС (ОК-5, ОПК5, ПК-1, 2, 3), критерий 5 АИОР (п.1.4)
P6	Внедрять, эксплуатировать и обслуживать современные программно-аппаратные комплексы, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды.	Требования ФГОС (ОК-8, 9, ОПК-1, 2, 4, ПК-3, 4, 5), критерий 5 АИОР (п. 1.5)
Универсальные компетенции		
P7	Использовать базовые и специальные знания в области проектного менеджмента для ведения комплексной инженерной деятельности.	Требования ФГОС (ОК-3, ОПК3, 5),

		критерий 5 АИОР (п. 2.1)
P8	Владеть иностранным языком на уровне, позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты комплексной инженерной деятельности.	Требования ФГОС (ОК-5, 7, ПК-3, 4), критерий 5 АИОР (п. 2.2)
P9	Эффективно работать индивидуально и в качестве члена группы, состоящей из специалистов различных направлений и квалификаций, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре организации.	Требования ФГОС (ОК-2, 6, 7), критерий 5 АИОР (п. 2.3, 2.4)
P10	Демонстрировать знания правовых, социальных, экономических и культурных аспектов комплексной инженерной деятельности.	Требования ФГОС (ОК-1, 2, 3, 4, 5), критерий 5 АИОР (п. 2.5)
P11	Демонстрировать способность к самостоятельному обучению в течение всей жизни и непрерывному самосовершенствованию в инженерной профессии.	Требования ФГОС (ОК-5, 7), критерий 5 АИОР (п. 2.6)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.03.01 «Информатика и вычислительная техника»
Отделение информационных технологий

УТВЕРЖДАЮ:

Руководитель ООП

_____ Погребной А.В.

(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы

Студенту:

Группа	ФИО
8B5A	Мнацаканян Рафаэль Эдуардович

Тема работы:

Предсказания повторных вызовов скорой помощи алгоритмами машинного обучения	
Утверждена приказом директора (дата, номер)	18.02.2019 № 1239/с

Срок сдачи студентом выполненной работы:	
--	--

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	В работе рассматривается применение методов машинного обучения для решения задач анализа медицинских данных. В том числе приводится описание и решение проблемы предсказания повторных вызовов скорой медицинской помощи пациентами с высоким риском развития сердечно-сосудистых катастроф.
Перечень подлежащих исследованию, проектированию и разработке вопросов	<ul style="list-style-type: none"> – Анализ предметной области, – Реализация решения реального времени, – Реализация подхода предотвращения кризиса, – Разработка модуля на основе разработанного алгоритма, – Финансовый менеджмент, – Социальная ответственность.
Перечень графического материала	
Консультанты по разделам выпускной квалификационной работы	
Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Шулинина Юлия Игоревна
Социальная ответственность	Мезенцева Ирина Леонидовна

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	
---	--

Задание выдал руководитель / консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Кудинов Антон Викторович	К.Т.Н		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8B5A	Мнацаканян Рафаэль Эдуардович		

Инженерная школа информационных технологий и робототехники
 Направление подготовки 09.03.01 «Информатика и вычислительная техника»
 Уровень образования – бакалавриат
 Отделение информационных технологий
 Период выполнения осенний / весенний 2018/2019 учебного года

Форма представления работы:

Бакалаврская работа

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	
--	--

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
01.12.2019	Анализ предметной области	10
12.03.2019	Реализация решения реального времени	35
01.06.2019	Реализация подхода предотвращения кризиса	35
05.06.2019	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
08.06.2019	Социальная ответственность	10

СОСТАВИЛ:

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Кудинов Антон Викторович	к.т.н		

СОГЛАСОВАНО:

Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Погребной Александр Владимирович	к.т.н		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8B5A	Мнацаканян Рафаэль Эдуардович

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	бакалавриат	Направление/специальность	09.03.01 Информатика и вычислительная техника

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

1. <i>Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих</i>	Оклад инженера – 21760 руб. Оклад руководителя – 33664 руб.
2. <i>Нормы и нормативы расходования ресурсов</i>	Премияльный коэффициент руководителя 30%; Коэффициент доплат и надбавок руководителя 20%; Районный коэффициент 30%. Коэффициент дополнительной заработной платы 12%; Накладные расходы 16%.
3. <i>Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования</i>	Коэффициент отчислений на уплату во внебюджетные фонды 30 %

Перечень вопросов, подлежащих исследованию, проектированию и разработке:

1. <i>Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения</i>	-Анализ конкурентных технических решений
2. <i>Планирование и формирование бюджета научных исследований</i>	Формирование плана и графика разработки: - определение структуры работ; - определение трудоемкости работ; - разработка графика Гантта. Формирование бюджета затрат на научное исследование: - материальные затраты; - заработная плата (основная и дополнительная); - отчисления на социальные цели; - накладные расходы.
3. <i>Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования</i>	- Определение эффективности исследования

Перечень графического материала (с точным указанием обязательных чертежей):

1. <i>Оценочная карта конкурентных технических решений</i>
2. <i>Матрица SWOT</i>
3. <i>График Гантта</i>
4. <i>Расчет бюджета затрат</i>

Дата выдачи задания для раздела по линейному графику	
---	--

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ОСГН ШБИП	Шулинина Ю.И.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5А	Мнацаканян Р.Э.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8В5А	Мнацаканян Рафаэль Эдуардович

Школа	ИШИТР	Отделение (НОЦ)	ОИТ
Уровень образования	бакалавриат	Направление/специальность	09.03.01 Информатика и вычислительная техника

Исходные данные к разделу «Социальная ответственность»:

1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Целью выпускной квалификационной работы является создание программного обеспечения, которое будет предсказывать вызовы скорой помощи. Сервис представляет собой десктопное приложение, следовательно, разработка и эксплуатация данного сервиса предполагает работу в помещении оператором ЭВМ.
--	---

Перечень вопросов, подлежащих исследованию, проектированию и разработке:

1. Правовые и организационные вопросы обеспечения безопасности: <ul style="list-style-type: none"> – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны. 	В разделе описаны правовые и организационные мероприятия по обеспечению рабочего места разработчика сервиса и пользователя сервиса. Организация рабочего места, оборудованного персональным компьютером, осуществляется в соответствии со следующими нормативными документами: ГОСТ 12.2.033-78 ССБТ, ГОСТ 12.2.032-78 ССБТ, СанПиН 2.2.2/2.4.1340-03.
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	В данном разделе представлен анализ опасных и вредных факторов, которые могут возникать при разработке или эксплуатации сервиса. Вредные факторы: <ul style="list-style-type: none"> • электромагнитное излучение; • повышенный уровень шума; • недостаточная освещённость рабочей зоны; • статические физические нагрузки; • перенапряжение зрительных анализаторов. Опасные факторы: <ul style="list-style-type: none"> • опасность поражения электрическим током; • опасность возникновения короткого замыкания; • повышенный уровень статического электричества.
3. Экологическая безопасность:	Негативное влияние на окружающую среду связано с утилизацией люминесцентных ламп, используемых для освещения рабочего помещения, а также с эксплуатацией персонального компьютера.
4. Безопасность в чрезвычайных ситуациях:	В данном разделе описана наиболее вероятная чрезвычайная ситуации при эксплуатации разработанного сервиса – возникновение пожара.

Дата выдачи задания для раздела по линейному графику	
--	--

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
ассистент ООТД	Мезенцева И.Л.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8В5А	Мнацаканян Рафаэль Эдуардович		

РЕФЕРАТ

Выпускная квалификационная работа: 72 с., 13 рис., 11 табл., 27 источников.

Ключевые слова: СМП, предсказание ухудшения состояния, сердечно-сосудистые катастрофы, врачебные приемы, машинное обучение, классификация, случайный лес, искусственная нейронная сеть, рекуррентные сети, LSTM, word embedding.

Объектом исследования являются данные о врачебных приемах и вызовах скорой за 2017-2018 года.

Целью работы является предсказание повторных вызовов скорой помощи алгоритмами машинного обучения.

В результате исследования проработаны несколько путей решения поставленной задачи. Разработаны классификаторы на основе математических классификаторов и на основе рекуррентных нейронных сетей. Модули показали частичную работоспособность, лучшую чем случайное гадание и значительно экономящие время персонала. До момента внедрения требуется улучшение точности алгоритмов и уменьшение количества ложных срабатываний.

Основные конструктивные, технологические и технико-эксплуатационные характеристики: модуль разработан на языке программирования Python и использующий библиотеки pandas, scikit-learn и keras.

Область применения: предсказания ухудшения состояния человека автоматизированными средствами.

Экономическая значимость работы заключается в автоматизации ухудшения состояния пациента, экономия времени работы врача и предотвращение тяжелых сердечно-сосудистых катастроф.

СОКРАЩЕНИЯ

СМП – скорая медицинская помощь

ССК – сердечно сосудистые катастрофы

ИНС – искусственные нейронные сети

МКБ – международная классификация болезней

Кардио пациенты – пациенты с проблемами по сердечно сосудистому профилю

ИС – информационные системы

ОГЛАВЛЕНИЕ

Оглавление	11
Введение	13
1. Анализ предметной области.....	13
1.1. Актуальность	13
1.2. Анализ проблематики	14
1.3. Существующие ресурсы	14
1.4. Формальная постановка задачи	16
1.5. Теоретическая основа классификации.....	16
1.5.1. Обзор используемых моделей классификации	17
1.5.1.1. Искусственные нейронные сети.....	17
1.5.1.2. Бустинг.....	18
1.5.1.3. Случайный лес	18
1.5.2. Используемые метрики для анализа результата.....	19
1.5.2.1. Матрица несоответствий.....	19
1.5.2.2. Метрики.....	19
1.6. Используемые программные инструменты	20
2. Реализация решения реального времени.....	21
2.1. Подготовка данных	21
2.2. Проведение обучения	25
2.3. Результат решения реального времени.....	26
3. Подход предотвращения кризиса	27
3.1. Общая архитектура	27
3.2. Препроцессинг текста	28
3.3. Векторизация слов.....	29
3.4. Модуль предоставления ответа.....	35
3.5. Выводы по данному решению	39
4. Финансовый менеджмент, ресурсоэффективность, и ресурсосбережение.....	40
4.1. Потенциальные потребители результатов исследования.....	40
4.2. Технология QuaD	41
4.3. SWOT-анализ.....	43
4.4. Планирование научно-исследовательских работ	44
4.4.1. Структура работ в рамках научного исследования	44
4.4.2. Определение трудоемкости выполнения работ	45
4.5. Бюджет научно-технического исследования	49

4.5.1.	Расчет затрат на специальное оборудование для научных работ	49
4.5.2.	Основная заработная плата исполнителей темы	51
4.5.3.	Дополнительная заработная плата исполнителей темы	53
4.5.4.	Отчисления во внебюджетные фонды	54
4.5.5.	Накладные расходы.....	55
4.5.6.	Формирование бюджета затрат научно-исследовательского проекта..	56
4.6.	Определение потенциального эффекта исследования	57
5.	Социальная ответственность	58
5.1.	Правовые и организационные вопросы обеспечения безопасности.....	58
5.1.1.	Специальные правовые нормы трудового законодательства	58
5.1.2.	Организационные мероприятия по компоновке рабочей зоны	59
5.2.	Производственная безопасность	60
5.2.1.	Вредные факторы	61
5.2.1.1.	Электромагнитное излучение.....	61
5.2.1.2.	Повышенный уровень шума	62
5.2.1.3.	Недостаточная освещенность рабочей зоны	63
5.2.1.4.	Статические физические нагрузки.....	63
5.2.1.5.	Перенапряжение глаз	64
5.2.2.	Опасные факторы	64
5.2.2.1.	Опасность поражения электрическим током.....	64
5.2.2.2.	Опасность возникновения короткого замыкания	65
5.2.2.3.	Повышенный уровень статического электричества	65
5.3.	Экологическая безопасность	66
5.4.	Безопасность в чрезвычайных ситуациях	66
5.5.	Выводы по разделу	67
Заключение	69
Список используемых источников	70

Введение

Социальная сфера жизни населения затрагивает абсолютно всех жителей нашей страны, поэтому важной целью является максимальное улучшение ее различных аспектов. В рамках запроса от заинтересованных лиц была выделена задача работы со скорой помощью и уменьшение издержек работы данной службы. Специфика работы данной отрасли сильно тормозит автоматизацию процессов современными информационными средствами, поэтому было решено провести исследование, возможна ли такая автоматизация и если возможна, то насколько эффективна и какие есть перспективы внедрения. В работе было проведено такое исследование, сделаны выводы и созданы исследовательские модели.

1. Анализ предметной области

1.1.Актуальность

При вызове скорой медицинской помощи люди часто отказываются от госпитализации, после чего их состояние может ухудшиться. Как результат – необходим повторный вызов медицинской помощи, что является тратой ограниченных ресурсов скорой, в результате чего происходит, как ухудшение качества реагирования на другие вызовы, так и прямая трата денежных средств.[1]

Согласно приказу Министерства здравоохранения и социального развития РФ от 2 декабря 2009 г. N 942 «Об утверждении статистического инструментария станции (отделения), больницы скорой медицинской помощи», повторный вызов – вызов к больному, который обращался на станцию (подстанцию, отделение) скорой медицинской помощи в течение последних 24 часов.[2]

По данным с января по июнь 2018 года диспетчерам скорой медицинской помощи по г. Томску поступило более 177 тысяч вызовов. Среди них 8% – повторные. Уменьшение их количества необходимо для

оптимизации функционирования СМП и уменьшения нагрузки на каждую бригаду, что положительно скажется как на качестве работы, так и на экономической эффективности.

Причин повторных вызовов может быть несколько, вот основные из них:

- неправильно или некачественно оказанная медицинская помощь;
- острое обострение заболевания;
- отказ пациента от госпитализации на первичном вызове с последующим ухудшением состояния.

Проанализировав природу данных причин, были созданы несколько возможных подходов к решению исходной проблемы.

1.2. Анализ проблематики

Основная проблематика задачи кроется в её природе, так как не существует автоматизированных систем оценки тяжести состояния пациента. Даже опытным врачам часто непросто справиться с этой задачей, поэтому первым приближением к решению задачи будет понимание, возможно ли принципиальное решение, или анализ данных не позволит с хорошей точностью предсказывать возникновение кризисных ситуаций у пациентов в будущем.

Для этого в нашем распоряжении есть два возможных подхода:

- Предсказание повторного вызова в реальном времени, сразу при первичном вызове;
- Предсказание кризисного состояния заранее, предотвращая проблему.

Попробуем оба подхода, создадим и обучим модели и оценим их метриками точности и построим матрицы несоответствий.

1.3. Существующие ресурсы

Для разработки были выделены обезличенные данные о врачебных приемах и о вызовах скорой помощи жителями города Томска за период 2017-

2018 годов. Данные были использованы для разработки моделей, которые планируются использоваться в работе в комплексе с существующими медицинскими ИС. После внедрения система будет автоматизировано решать задачу классификации тяжести состояния пациента.

Данные были переконвертированы из формата Microsoft database (.mdb) в свободно распространяемый формат электронных таблиц (.csv), более удобный для дальнейшей обработки автоматизированными средствами.

Каждый вызов скорой представлен одной записью, состоящую из 27 полей. Вот некоторые из них:

- Пол
- Дата рождения
- Дата вызова
- Время вызова
- Основной диагноз в кодификации МКБ10
- Дом
- Улица
- Вид заболевания (острое, хроническое и др.)
- Категория вызова (первичная или повторная)
- Результат выезда
- Идентификатор человека
- Идентификатор вызова скорой помощи

Все поля в данной таблице являются описывающие какую-либо характеристику вызова в одно слово, временную метку, код или букву. Общее количество вызовов – 285009.

Каждый врачебный прием так же представлен одной записью уже в другой таблице. Каждая такая запись состоит из 11 полей:

- Идентификатор человека
- Дата приема

- Тип приема и специализация врача
- Основной диагноз в кодификации МКБ10
- Жалобы
- Анамнез
- Объективный статус
- Рекомендации
- Код врачебного учреждения
- Дата первого вызова скорой этим пациентом
- Идентификатор вызова скорой помощи

Поля жалобы, анамнез, объективный статус и рекомендации являются текстовыми, в которых хранятся записи врача, сделанные по ходу приема в свободной форме

1.4.Формальная постановка задачи

Нашей задачей является обрабатывать каждый вызов, как набор сигналов, и отдавать в результате ответ, есть ли риск повторного вызова, или нет. Фактически задача сводится к бинарной классификации (есть риск или нет), позволяя применить весь существующий математический аппарат по данной задаче для её решения.

1.5.Теоретическая основа классификации

Классификация – раздел машинного обучения, в котором решается задача сортировки объектов по группам на основании специфических признаков. В нашем случае у каждого объекта есть добавленная метка, указывающая принадлежность классу, таким образом решаемая задача – классификация с учителем.[3]

Для решения задачи формируются тесты, которые разделяются на две группы, тренировочную и валидационную выборки. Тренировочная выборка используется непосредственно во время обучающих прогонов классификатора, во время которых вычисляются значения функции потерь и

веса изменяются в попытке её уменьшить. Так как обучение классификатора — это итерационный процесс, контролируемый инженером по машинному обучению, необходимо понять, когда дальнейшее обучение классификатора приведет лишь к заучиванию тестового примера.

Это известная проблема, которая называется переобучение (англ. *overfit*), классификатор будет работать на тестовой выборке все лучше и лучше, в ущерб общности предсказания, которые станут не стабильными при работе с данными, которые не были включены в тестовую выборку. Для решения данной проблемы используется валидационная выборка, на которой высчитывают значение функции ошибок на момент завершения каждой обучающей эпохи. Как только ошибки на данных, отсутствующих в тренировочной части начинают увеличиваться – обучение следует остановить.

Самих моделей классификации, обучаемых в соответствии описанными выше правилами достаточно много. Они условно поделены по применимости в разных областях знаний, однако чаще всего сложно однозначно утверждать какой классификатор окажется наилучшим, поэтому в работе были проверены несколько популярных классификаторов.

1.5.1. Обзор используемых моделей классификации

1.5.1.1. Искусственные нейронные сети

Искусственными нейронными сетями называют широкий спектр математических моделей, которые были изобретены под влиянием исследования нервной системы живых существ и применения к этим исследованиям математического аппарата. Не смотря на разницу в исполнении видов нейронных сетей, всех их можно описать как связанный взвешенный граф, хранящий простые процессоры в вершинах, а пропуская входящий сигнал от истока графа к стоку система и выдает, полученный в результате множества элементарных математических операций, выход, являющийся ответом системы.[5]

Веса связей в ИНС и являются настраиваемыми параметрами, которые позволяют решать огромное количество различных задач, от работы с изображениями и видео, до анализа простых комбинационных данных. Поскольку разные слои ИНС часто решают абсолютно различные задачи и имеют свою специфику, подробности работы каждого применяемого слоя будут описаны в последующих главах.

1.5.1.2. Бустинг

Бустингом называют объединение множества элементарных моделей, самостоятельная классификация которыми фактически эквивалентна случайным предсказаниям. За время обучения устанавливаются веса и происходит отбор нужных случайных классификаторов. Таким образом множество использованных случайных классификаторов в результате дает приемлемый по точности результат.[11]

Существует много видов бустинга, основным различием в которых является именно природа изначального элементарного алгоритма. В работе были использованы градиентный буст (англ. gradient boost) и АДА буст(англ, adaboost).

1.5.1.3. Случайный лес

Случайный лес – это модель, основанная на ансамбле деревьев решений. Само по себе дерево является набором правил, которое по ответам на вопросы позволяет классифицировать некоторые объекты.[8] Точность предсказания единственного дерева невелика по причине ограничения его глубины, в свою очередь несколько деревьев, решающие одну и ту же задачу, итоговый результат которой выбирается общим голосованием позволяет значительно улучшить результат.[6]

1.5.2. Используемые метрики для анализа результата

1.5.2.1. Матрица несоответствий

Матрица несоответствий (англ. confusion matrix) – матрица, в которой для каждого класса наблюдений приводится количество наблюдений, отнесенных моделью к этому и другим классам. Каждая строка матрицы соответствует каждому классу, по которым идет разделение обученным классификатором. В каждой ячейке такой матрицы содержится число – количество предсказанных классов. Если данная ячейка содержится на главной диагонали, то класс был предсказан верно, иначе строка в которой находится данное значение – это истинный класс, а столбец – тот, который был предсказан неправильно. Таким образом матрица позволяет не только увидеть насколько правильно работает классификатор, но и точно понять, какие классы плохо различимы. Ее общий вид представлен на Рисунке 1.

		<i>P</i>	<i>N</i>
True label	<i>P</i>	True Positives (TP)	False Negatives (FN)
	<i>N</i>	False Positives (FP)	True Negatives (TN)
		Predicted label	

Рисунок 1 – Общий вид матрицы несоответствий

Преобладание на главной диагонали матрицы несоответствий означает, что исследуемый метод устойчив и классы хорошо разделимы.

1.5.2.2. Метрики

Ассигасу – доля правильных ответов алгоритма. Вычисляется по следующей формуле: $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.

Precision характеризует способность модели отличать данный класс от других классов. Вычисляется по формуле: $precision = \frac{TP}{TP+FP}$.

Recall демонстрирует способность модели обнаруживать класс. Вычисляется по формуле: $recall = \frac{TP}{TP+FN}$.

F-мера – способ объединить precision и recall. Чем больше f-мера, тем более явно выражено диагональное преобладание в матрице несоответствий. В работе использовался распространенный способ вычисления f-меры – f1-score. Формула: $f1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$.

Какие метрики использовать для выбора лучшей модели, зависит от постановки задачи и требуемого результата.[15]

1.6.Используемые программные инструменты

В работе использовался стандартный набор инструментов для анализа данных и машинного обучения.

Язык программирования Python – высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода.[12]

Python отличается богатой стандартной библиотекой, лаконичным синтаксисом. Python является одним из основных языков программирования для анализа данных и машинного обучения. Он прост в изучении, доступен практически на всех платформах, поддерживает использование в интерактивном режиме. Для Python существует множество библиотек.

Основные из использованных в работе библиотек:

- scikit-learn – библиотека для языка Python, содержащая большое количество алгоритмов машинного обучения, в том числе реализации вышеупомянутых классификаторов;
- Keras, в связке с TensorFlow в качестве backend – библиотека для языка Python для работы с искусственными нейронными сетями и средствами их обучения;

- Pandas – библиотека для языка Python, содержащая инструменты для работы с данными.

2. Реализация решения реального времени

Для реализации подобного подхода необходимо использовать данные вызовов скорой помощи, так как во время реальной работы у СМП в наличии будет только данные о вызове в данный момент, по которому они уже будут решать, требуется ли госпитализация или нет. Система должна решать, будет ли повторный вызов, если человека не госпитализировать сейчас.

Для анализа вызова доступно только информация о текущем вызове из таблицы вызовов скорой. Проблема состоит в том, что для честного анализа нам не доступна информация о пациенте, поскольку диспетчер СМП часто не обладает всей информацией для немедленной связи текущего вызова с историей болезни пациента.

2.1. Подготовка данных

В первую очередь для любого алгоритма классификации необходимо подготовить данные с целью искоренения грубых ошибок, стандартизации и нормализации цифровых значений.

Первичная подготовка данных включала в себя:

- Добавление расшифровок закодированных полей для более простой интерпретации данных. Сопровождающая документация к данным была не полной, из-за чего потребовалась идентификация содержащихся данных.
- Данные были отчищены от вызовов, после которых не последовали выезды скорой помощи, случаи, когда бригада не встретила с пациентом по тем или иным причинам, вызовы с малым количеством заполненной информации, а также телефонные консультации.

После первичной обработки данных были выделены следующие информативные признаки:

- Пол;
- Дата рождения;
- Дата вызова;
- Время вызова;
- Основной диагноз;
- Дом;
- Улица;
- Вид заболевания;
- Категория вызова;
- Результат выезда;
- Вызывающий;
- АО.

Некоторые признаки потребовали трансформации для дальнейшего использования. Вызывающие были разбиты на три категории «БЛИЗКИЙ», «САМ», «КТО-ТО» путем составления словаря всех возможных вариантов родственников и их опечаток, а также «САМ»-«САМА»-«ПОСТРАДАВШИЙ» и всех возможных опечаток.

Дома были приведены к виду, где сохранялись числа, разделенные символом дроби и одиночные буквы (предположительно, именованные корпусов).

Улицы прошли несколько вариантов трансформации, но большое количество опечаток, различных стилей записи и сокращений в этом поле вынудили вместо классических решений (например, применения готового списка улиц города и расстояния Левенштейна) применять более тонкий подход. Последняя версия трансформации улиц удаляла все небуквенные символы, сокращала слова и склеивала их вместе.

Данные были очищены от людей нереального возраста и младенцев в алкогольном опьянении (положительное значение поля «АО» при возрасте пациента до года).

Так же были добавлены следующие косвенные признаки:

- Время ожидания – предположительно, показатель важности вызова;
- Время в пути – географическая удаленность человека (или наличие дорожных пробок);
- Время на месте – насколько тяжелая ситуация;
- День в году – предположительное влияние сезонности;
- День недели – предположительное влияние выходных;
- Тип бригады: линейная, кардиологическая, реанимационная;
- Температура воздуха – вероятное влияние погодных условий;
- Абсолютное атмосферное давление – вероятное влияние погодных условий;
- Изменение атмосферного давления за 3 часа – вероятное влияние погодных условий.

По результатам проведенного при помощи метода главных компонент (преобразование Карунена – Лозва) предварительного анализа, признаки не всегда оказывали положительное влияние на предсказания повторных вызовов.

Положительное влияние:

- День в году;
- Время на адресе (особенно релевантный признак для классификаторов);
- Пол;
- Возраст;
- Основной диагноз;

- Вид заболевания;
- Атмосферное давление;
- Изменение давления.

Отрицательное влияние:

- Минута дня;
- Вызывающий;
- Время в пути (особенно нерелевантный признак для

классификаторов);

- Тип бригады;
- АО.

Неоднозначное влияние:

- День недели;
- Температура.

Общение с заинтересованными в разработке лицами, выдавшими техническое задание, позволило скорректировать задачу в распознавание повторных вызовов у пациентов, имеющих проблемы по кардиологическому профилю. Здесь и далее будет решаться именно эта задача.

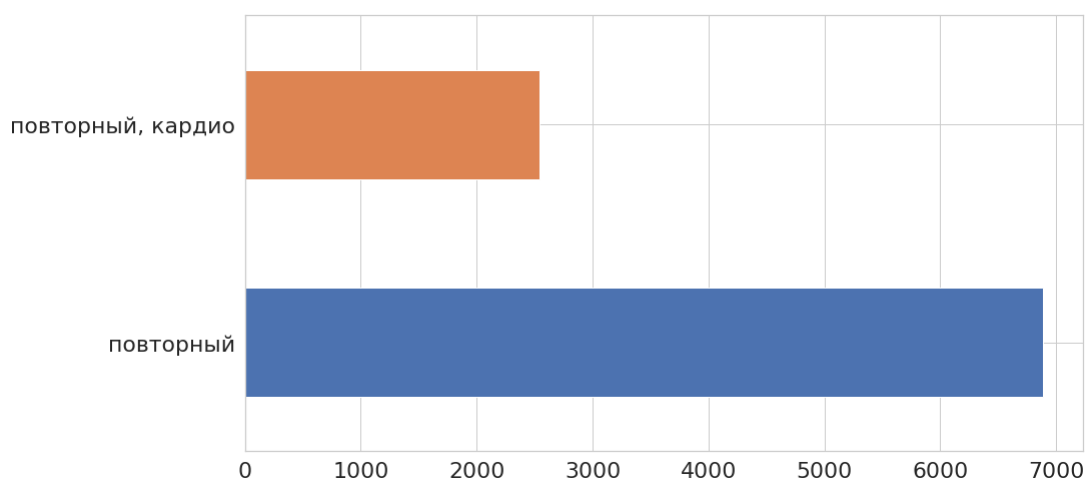


Рисунок 2 – Пропорциональное распределение всех повторных вызовов и повторных вызовов людей, имеющих проблемы по кардиологическому профилю.

Обучающей выборкой для нас были кардио-пациенты, отказавшиеся от госпитализации на первичном вызове.

Для каждого повторного вызова ставилась пометка на его первичном вызове, которая и являлась ответом для математических алгоритмов. При наличии повторного вызова – 1, при отсутствии – 0.

2.2.Проведение обучения

Было проведено несколько попыток обучения на полученных данных. Использованы модели случайного леса, ада буста и градиентного буста и прочносвязанных сетей разных архитектур.

К сожалению, диагонального преобладания нет ни при одном классификаторе, или нейронной сети. Важных прямых данных мало, а угадывание по косвенным признакам оказалось неточным. Тем не менее, по итогу даже при этих условиях можно выделить лучший результат при данном подходе к проблеме.

Для решения проблемы необходимо максимизировать точность, так важно спасти как можно больше пациентов, затрачивая как можно меньше ресурсов. Далее приведены лучшие метрики, полученные на наилучшем классификаторе, основанном на случайном лесе.

ACCURACY: 0.783

F1-SCORE: 0.574

PRECISION: 0.589

RECALL: 0.567

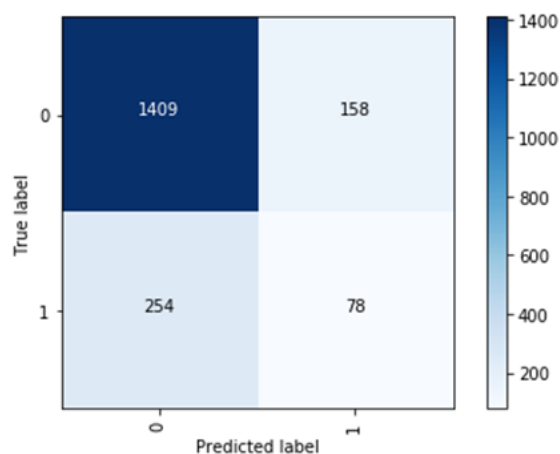


Рисунок 3 – Матрица несоответствия, по которой предсказываются 78 повторных вызовов из 332 (23,5%), однако классификатор 158 раз срабатывает ошибочно, предсказывая повторный вызов там, где не следовало

2.3.Результат решения реального времени

- Лучшее полученное решение помогает избежать 23,5% повторных вызовов скорой медицинской помощи
- Имеющихся признаков недостаточно, чтобы достаточно точно определять необходимость госпитализации
- Необходимы данные медицинских карт пациентов для более точного понимания состояния здоровья в момент работы скорой помощи, что позволит более точно предсказывать вероятность повторного вызова.
- Для улучшения работы и расширения функционала классификатора, необходимы данные за полный год, для того, чтобы точнее проанализировать зависимость вызовов от времени года, даты, возможных эпидемий, и погодных условий.
- Возможно, необходимо иметь описание состава врачебной бригады для понимания, корреляции качества оказания помощи врачей разного пола и возраста, разным категориям граждан

Таким образом подход слабо перспективен из-за низкого содержания медицинских данных в процессе предсказывания, попробуем другой подход.

3. Подход предотвращения кризиса

3.1.Общая архитектура

Анализируя данные врачебных приемов (описаны в пункте 1.3), была обнаружена их низкая классифицируемость и общность. Данные описывают врачебный прием, содержа большое количество текстового описания состояния пациента. Фактически интересными полями для нас являются поля жалобы, анамнез, объективный статус и рекомендации. С помощью этих полей решается фундаментальная проблема предыдущего метода, малое количество медицинского смысла в предсказаниях. В свою очередь это привело к иной проблеме – автоматизированному семантическому анализу текста. Данная задача является одним из самых популярных объектов современных исследований в машинном обучении. Для работы мы применим сформировавшийся подход к решению данных проблем.

Ручные данные содержали множество разночтений и не имели общего стиля, из-за чего было решено применить именно алгоритмы машинного обучения, основанные на нейронных сетях. Выбор был обусловлен неспособностью математических классификаторов точно работать с библиотеками слов и с временными рядами.

Сам программный продукт условно разбит на три модуля.

Первым в архитектуре находится блок текстовой предварительной подготовки. Он необходим для стандартизации текста. Модуль удаляет все небуквенные символы и разбивает строку ввода на подстроки – слова, разделённых пробелами, или какими-либо знаками пунктуации. Эти слова переводятся в вид с исключительно строчными буквами, все окончания отбрасываются. Слова, состоящие из менее чем 2 символов, отбрасываются. Далее каждое слово проверяется на наличие в предварительно подготовленном словаре из 4000 слов, слова вне этого списка отбрасываются по причинам аппаратных ограничений размера памяти в графических ускорителях,

используемых в обучении. На выходе блок предоставляет обработанную строку из слов, разделённых пробелом.

Вторым блоком находится предварительно подготовленный модуль, состоящий из специальной нейронной сети. Данная сеть предварительно обучена, она получает слова по одному и для каждого генерирует вектор размером 1×256 . Данные численные вектора близки по значению, если слова семантически близки по смыслу. Выходом данного модуля является вектор размера $n \times 256$, где n - количество слов после обработки первым блоком.

Располагающиеся последним блок – может так же предварительно обученные нейросети, которые в свою очередь решает задачу классификации. Используемая рекуррентная нейросеть получает на вход вектор-слово, выдает ответ на основании предыдущих слов, полученных ранее. Благодаря этому ответ контекстно зависим, что улучшает точность классификатора, так и его соответствие представлениям реального мира. Выходом данного блока является предположение класса, который является временным промежутком от данного приема до вызова человеком СМП по сердечно сосудистым причинам.

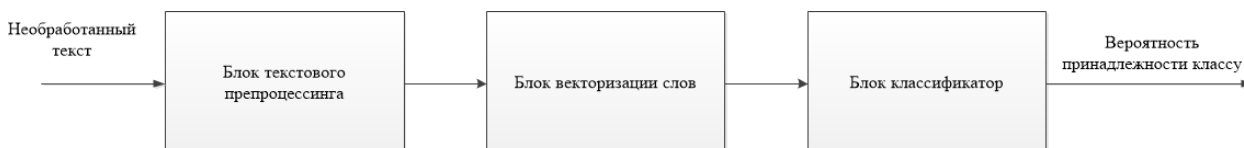


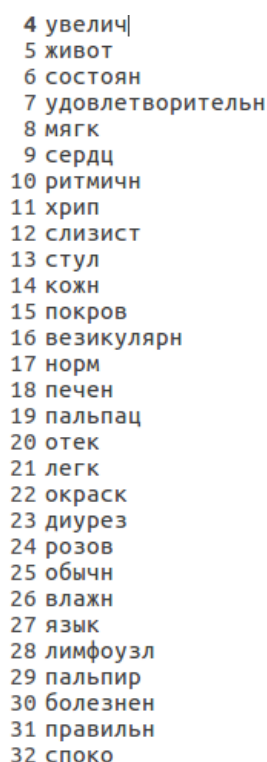
Рисунок 4 – Общая схема программного модуля.

3.2.Препроцессинг текста

Модуль реализует последовательную обработку массива текста. Для лучшей работы классификаторов необходимо иметь наиболее релевантные входные данные. С этой целью производится работа, по максимальному улучшению текстовой части. Для анализа текста по сочетанию ключевых слов нет необходимости хранить пунктуацию. Слова часто содержатся в форме различных лексем. Семантически форма слова важна, но для анализа гораздо

важнее понимать, что какое-то слово может применяться в двух разных контекстах, чем-то что у нас есть два разных слова, которые применяются по-разному.

Таким образом, работа модуля заключается в черновой подготовке текста, из которого потом выбираются 4000 слов, находящиеся в предварительно подготовленной библиотеке. Каждый врачебный прием рассматривается как текстовая строка. Строка токенизируется и сохраняется в лист строк-слов. Данные слова очищаются от специальных символов и цифр, после чего стеммируются (лишаются окончаний). По результатам проверки удаляются очень короткие слова и слова из стоп листа. Слова заменяются на индекс из словаря для эффективного хранения большого объема тестовых данных.



4 увелич|
5 живот
6 состоян
7 удовлетворительн
8 мягк
9 сердц
10 ритмичн
11 хрип
12 слизист
13 стул
14 кожн
15 покров
16 везикулярн
17 норм
18 печен
19 пальпац
20 отек
21 легк
22 окраск
23 диурез
24 розов
25 обычн
26 влажн
27 язык
28 лимфоузл
29 пальпир
30 болезнен
31 правильн
32 споко

Рисунок 5 – Участок словаря с индексами слов.

3.3. Векторизация слов

Проблемой, стоящей перед нами, является отсутствие понимания алгоритмами машинного обучения семантической нагрузки слов. Так как

архитектурно невозможно их использование как текста используются их индексы, что ведет к тому, что любые два слова будут абсолютно одинаково восприниматься нейросетевыми алгоритмами, которые будут смотреть на два этих числа с точки зрения каких-то различных токенов без привязки их к какому-либо смысловому значению.

К тому же при переводе обработанных в предыдущем модуле слов-индексов к виду one-hot кодирования, каждое слово будет занимать огромное количество пространства, большая часть которого будет заполнена нулями. Для преодоления данной ситуации применяется векторизация слов, которая превращает разреженный вектор one-hot кодирования в плотный вектор хеширования или любого другого алгоритма компрессии.



Рисунок 6 – Визуальное сравнение векторов прямого кодирования и вектора представления.

Плотная компрессия, однако, не решает проблему отсутствия семантической нагрузки. Вместо одной случайные единицы в массиве, здесь находится несколько случайных чисел, распределение которых зависит только от выбранного алгоритма компрессии. Для решения этой задачи и улучшения

работы дальнейших алгоритмов, чтобы сети извлекали релевантную медицинскую информацию, а не гадали на сугубо статистических данных, был создан модуль реализующий специфическую векторизацию слов при помощи отдельного нейросетевого алгоритма.

Для работы данного модуля была проведена работа по подбору архитектуры нейронной сети и её обучение с различными значениями настроек скорости обучения и размера упаковок тестов.

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 7, 256)	1024256
flatten_1 (Flatten)	(None, 1792)	0
dropout_1 (Dropout)	(None, 1792)	0
dense_1 (Dense)	(None, 1024)	1836032
leaky_re_lu_1 (LeakyReLU)	(None, 1024)	0
dense_2 (Dense)	(None, 1024)	1049600
leaky_re_lu_2 (LeakyReLU)	(None, 1024)	0
dense_3 (Dense)	(None, 35)	35875
Total params: 3,945,763		

Рисунок 7 – Архитектура сети.

Данная сеть состоит из 6 слоев. Первый слой – специальный embedding слой, который создан именно для решения задачи векторизации слов. Впоследствии только этот слой используется как основа второго модуля.

Второй слой – flatten необходим для разворачивания матрицы выхода embedding слоя в вектор строку, для того, чтобы можно было использовать дальнейшую обработку данных последующими слоями.

Третий слой – dropout, необходим для предотвращения переобучения модели. Он случайно отключает 50% нейронов предыдущего слоя, запрещая кооперацию нейронов для улучшения результата в ущерб семантической логичности модели.

Три оставшийся слоя – классические полносвязные слои, которые обучаются и в результате выдают вектор из 35 чисел. Данный вектор указывает на то, какому классу соответствует кортеж из 7 введенных слов.

Обучение производится на подготовленных тестах, где вход – 7 слов, а выход – то, какое заболевание по кодификации МКБ-10 было найдено у этого пациента в процессе врачебного приема. Обучение производится на 21045577 тестовых примерах, которые были предварительно нами сформированы. Примеры формировались по принципу скользящего окна, которое захватывало 7 стоящих рядом слов. Такая ширина окна обусловлена проведенными экспериментами. Меньшие размеры не обеспечивают точности классификатора, что ухудшает точность семантических связей, формируемых embedding векторов, а большие размеры лучше для классификации, но из-за эффекта размывания сигнала, по многим словам, качество векторизации одного конкретного слова падает.

В результате работы, модуль реализует преобразование слова в векторы таким образом, что близкие по смыслу слова кодируются близкими по смыслу векторами. Таким образом будет достигнуто улучшение устойчивости ответа и ликвидация инвариантности от незначительной переформулировки фразы. Лучший результат был достигнут на 656 эпохе сети, которая классифицировала 7 граммы по 35 группам диагнозов МКБ10. Это заняло более 220 часов вычислений на видеокарте Nvidia Gforce 1060 6GB с постоянным уровнем утилизации более 70%. Значение функции потерь категориальной кросс-энтропии на данных анамнеза – 0.9968. Значение точности на данных анамнеза пациента — 70%, на данных объективного статуса — до 60 процентов.

Лучший промежуточный классификатор позволит максимально улучшить наши данные для дальнейших этапов, по этой причине проанализируем результаты работы, метрики (таблица 1 и таблица 2) и матрица несоответствия (рисунок 8).

Таблица 1 – Значения метрик по классам

Название класса	precision	recall	f1-score	support
Острая ревматическая лихорадка	0.53	0.29	0.37	90
Хронические ревматические болезни сердца	0.50	0.24	0.33	2278
Болезни характеризующиеся повышенным давлением	0.26	0.33	0.29	188201
Ишемическая болезнь сердца	0.37	0.38	0.38	154118
Легочное сердце	0.42	0.04	0.07	891
Другие болезни сердца	0.43	0.11	0.18	21722
Цереброваскулярные заболевания мозга	0.45	0.41	0.43	207062
Болезни артерий, артериол и капилляров	0.51	0.39	0.44	22223
Болезни вен, лимфотических сосудов и узлов	0.50	0.38	0.43	31783
Другие и неуточненные болезни системы кровообращения	0.62	0.12	0.19	1485
Врожденная аномалия системы кровообращения	0.54	0.38	0.44	9808
Травма сердца	0.27	0.13	0.17	31
Сердечно-сосудистые новообразования	0.45	0.47	0.46	15819
Симптомы и признаки системы кровообращения	0.46	0.41	0.44	13856
Инфекционно-паразитарные заболевания 1	0.51	0.34	0.41	35918
Инфекционно-паразитарные заболевания 2	0.45	0.28	0.35	38477
Новообразования (исключая СС)	0.64	0.66	0.65	231770
Болезни крови	0.50	0.18	0.26	139182
Болезни эндокринной системы	0.68	0.64	0.66	175059
Психические проблемы	0.83	0.66	0.73	12836
Болезни нервной системы	0.45	0.23	0.30	149703
Болезни глаза	0.69	0.60	0.64	30926
Болезни уха	0.54	0.35	0.43	64393
Болезни органов дыхания	0.57	0.72	0.64	827193
Болезни органов пищеварения	0.63	0.63	0.63	396019
Болезни кожи и подкожной клетчатки	0.51	0.36	0.42	64594
Болезни костно мышечной и соединительной ткани	0.52	0.54	0.53	350635
Болезни мочеполовой системы	0.54	0.60	0.57	293830
Беременность, роды и послеродовой период	0.61	0.57	0.59	118673
Отдельные состояния перинатального периода	0.85	0.83	0.84	22163
Пороки развития (исключая систему кровообращения)	0.57	0.10	0.18	9260
Выявленные симптомы неучитанные выше	0.35	0.13	0.19	37716
Травмы, отравления	0.55	0.52	0.54	130997
Факторы, влияющие на здоровья населения	0.55	0.59	0.57	801531
Внешние причины и другое	0.52	0.23	0.32	39333

Таблица 2 – среднее значение метрик

	precision	recall	f1-score	support
Классовое среднее	0.54	0.54	0.54	4639575
Общее среднее	0.52	0.40	0.43	4639575
Взвешенное среднее	0.54	0.54	0.53	4639575

Тема (row)	Остаток репутационной лиганды																																						-0.75
	0.29	0.011	0.019	0.23	0	0.0022	0.011	0	0	0	0	0	0	0	0	0	0	0	0.0011	0	0	0	0	0	0	0	0	0	0.014	0.022	0	0	0.011	0.011	0	0	0	0.033	0
Хронометрическая репутационная боковая серия -	0	0.24	0.013	0.29	0	0.0097	0.043	0	0.00088	0	0.0022	0	0.0013	0	0	0.0013	0.013	0.0031	0.019	0	0.0026	0.0013	0.00044	0.014	0.025	0.0013	0.025	0.0097	0.0031	0	0	0	0	0	0.0079	0.013	0.0026		
Боковая серия репутационных логических данных -	0	0.0001	0.33	0.12	4.3e-05	0.00318	0.046	0.0017	0.0022	0.0011	0.00013	0	0.00019	0.00018	0.00013	0.0015	0.0017	0.0026	0.024	0.000318	0.0052	0.0026	0.0016	0.022	0.054	0.0027	0.061	0.015	0.0088	0.0103	5.3e-06	0.00041	0.015	0.0061	0.0027				
Условная боковая серия -	1.9e-05	0.00064	0.016	0.38	3.9e-05	0.006	0.042	0.0025	0.0027	0.0027	0.00093	0	0.002	6.5e-06	0.00018	0.00088	0.002	0.0022	0.023	0.0004	0.0023	0.0027	0.0012	0.016	0.043	0.0025	0.041	0.013	0.012	3.9e-05	0	0.00041	0.02	0.0027	0.0016				
Логическая серия -	0	0.00022	0.022	0.23	0.004	0.015	0.015	0	0.0111	0	0	0	0.009	0.0022	0.0011	0.0011	0.008	0.0034	0.012	0.0022	0.0022	0	0	0.026	0.022	0	0.028	0.016	0.016	0	0	0	0	0.03	0.033	0.0022			
Другие боковые серии -	0	0.0006	0.012	0.24	0.00041	0.011	0.024	0.00078	0.0025	0	0.0024	0	0.0013	0.042	0.0005	0.0011	0.016	0.0022	0.019	0.00041	0.015	0.0015	0.0011	0.013	0.037	0.0016	0.03	0.012	0.01	0.0018	2e-05	0.00050	0.007	0.013	0.004	0.0027			
Церковносудовое заболевание когма -	1.4e-05	0.02e-05	0.099	0.042	4.8e-06	0.00027	0.41	0.002	0.012	2.9e-05	0.31e-05	0	0.00112	2.9e-05	0.00010	0.00071	0.014	0.002	0.012	0.001	0.043	0.0022	0.0017	0.09	0.03	0.0022	0.10	0.0093	0.0029	0.3e-05	0.0022	0.007	0.006	0.014					
Боковые артерий, артерий и капилляров -	0	0	0.041	0.038	0	0.0016	0.029	0.35	0.076	0	0	0	0.00076	0	0.0013	0.00089	0.026	0.0072	0.017	0.00035	0.0043	0.0022	0.001	0.055	0.12	0.022	0.091	0.0098	0.0064	4.5e-06	0.00018	0.026	0.007	0.0074					
Боковые вен, капиллярных сосудов и улик -	0	0.31e-05	0.017	0.028	0	0.0005	0.016	0.048	0.38	0.31e-05	6e-05	0	0.00076	6e-06	0.00037	0.012	0.0006	0.01	0.016	0.00025	0.0025	0.0019	0.0013	0.056	0.11	0.02	0.11	0.044	0.0001	0	0.00013	0.001	0.034	0.008	0.0053				
Другие и неопределенные боковые системы кровообращения -	0	0.002	0.15	0.11	0.00067	0.004	0.007	0.00000	0.001	0.12	0	0	0.0014	0	0.002	0	0.021	0.00027	0.011	0.002	0.037	0.01	0.002	0.24	0.037	0.0027	0.026	0.011	0.035	0.013	0	0.002	0.044	0.008	0.010				
Промежуточная логическая система кровообращения -	0</																																						

Хорошо заметно диагональное преобладание, что подтверждается приведенными выше метриками. Плохо различимые классы в левом верхнем

углу матрицы принадлежат сердечно сосудистым проблемам, что позволяет заключить что приведенная модель соответствует представлениям реального мира.

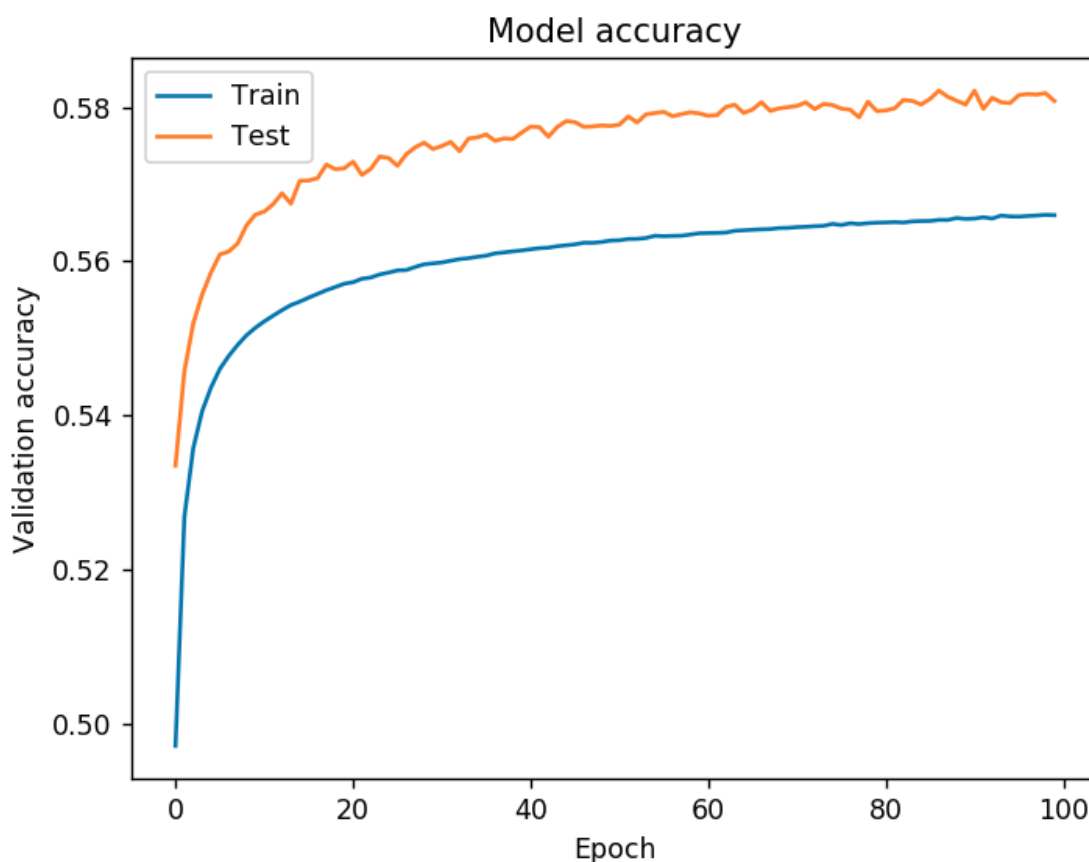


Рисунок 9 – Отображение зависимости значения точности от эпохи обучения на первой сотне эпох.

3.4. Модуль предоставления ответа

Классические нейросетевые алгоритмы прямого распространения не умеют работать с временными рядами. Если все данные загрузить за один проход работы, то сети смогут их успешно обработать, однако если данных много, то данная сеть станет громоздкой и её обучение может занять непозволительно много времени. Для преодоления данной проблемы были придуманы рекуррентные сети, которые хранят предыдущее состояние при загрузке новых данных. Данное состояние каждый раз изменяется под

действием функции обратного распространения ошибки, что поддерживает её актуальность.

Проблема возникает при обучении сложных моделей, в которых нужные связи разбросаны на более чем 2-3 эпохи. Состояние рекуррентной сети подвержено градиентному размытию, из-за чего долгосрочные связи просто не сохраняются для корректной обработки. Специально для этого изобретена принципиально улучшенная схема, которая так же хранит не размытые предыдущие состояния на виртуальном конвейере, который может предоставить историю нейрону в нужный момент. Данные сети называются LSTM сетями или сети с долгой краткосрочной памятью.

Основной идеей данных модулей является именно использование LSTM сетей, которые способны формировать контекстно-зависимое восприятие вводимых данных.

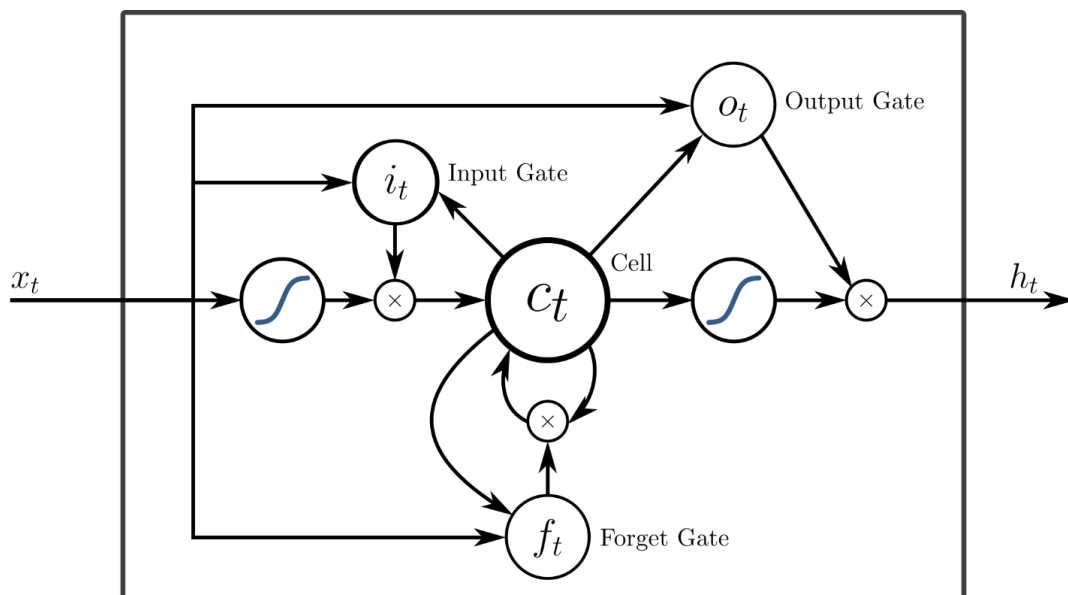


Рисунок 10 – Основа LSTM сети ячейка с долгой краткосрочной памятью.

Данные ячейки хранят в себе некоторое количество предыдущих вводов, не размывая их функцией активации или функцией обратного распространения ошибки. Благодаря этому, ячейки будут формировать свой выход, основываясь на нескольких словах, фактически учитывая контекст и приближаясь к человеческому типу понимания текста. Выходом данного

модуля является собственно вектор классов, где каждый класс соответствует срочности вызова скорой после вызова каким-либо человеком для классификатора.

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(16, 7, 256)	1024256
lstm_1 (LSTM)	(16, 7, 256)	525312
lstm_2 (LSTM)	(16, 128)	197120
dense_1 (Dense)	(16, 64)	8256
dense_2 (Dense)	(16, 6)	390
Total params: 1,755,334		
Trainable params: 731,078		
Non-trainable params: 1,024,256		

Рисунок 11 – Архитектура итоговой сети классификатора.

Первым слоем сети является embedding слой, обученный нами ранее. Он обрабатывает индексы слов на лету и превращает их в плотные вектора. Данный слой не обучается.

Два следующих слоя — это stateful (устойчивые) LSTM сети. Основным отличием таких сетей от неустойчивых является то, что мы вручную можем сбрасывать состояние нейронов, при помощи чего будем контролировать какие фразы связаны, а какие нет. Так как нам очень важно проследить развитие состояние пациента, мы загружаем его данные о приемах последовательно, соблюдая хронологический порядок. После окончания всех данных одного пациента, состояние вручную сбрасывается, для того чтобы разные пациенты не считались сетью одним человеком.

Располагающиеся далее два плотных слоя позволяют преобразовать специфический выход предыдущих сетей при помощи классических полносвязанных сетей в выход классификатор.[27]

Во время обучения были обнаружены значительные проблемы в преобладании классов, важных для заказчика, классами, которые важны для отсеивания ненужных приемов, но не несущих полезной информации.

Все проводимые эксперименты по обучению на данный момент имеют достаточно низкие значения gsc и f меры, что не позволит применять решения на данном этапе из-за большого количества ложных срабатываний, что сводит на нет эффект от верно предсказанных сердечно сосудистых катастроф. Проблема так же состоит в длительности обучения. Даже высокоскоростные GPU ускорители тратили порядка нескольких суток на полное обучение сети. Для нашей LSTM сети одна эпоха обучения занимает более 3 часов, что не позволяет провести большое количество экспериментов в сжатые сроки. Тем не менее лучший результат хоть и не имеет диагонального преобладания, позволяет установить некоторое количество требуемых ВЫЗОВОВ.

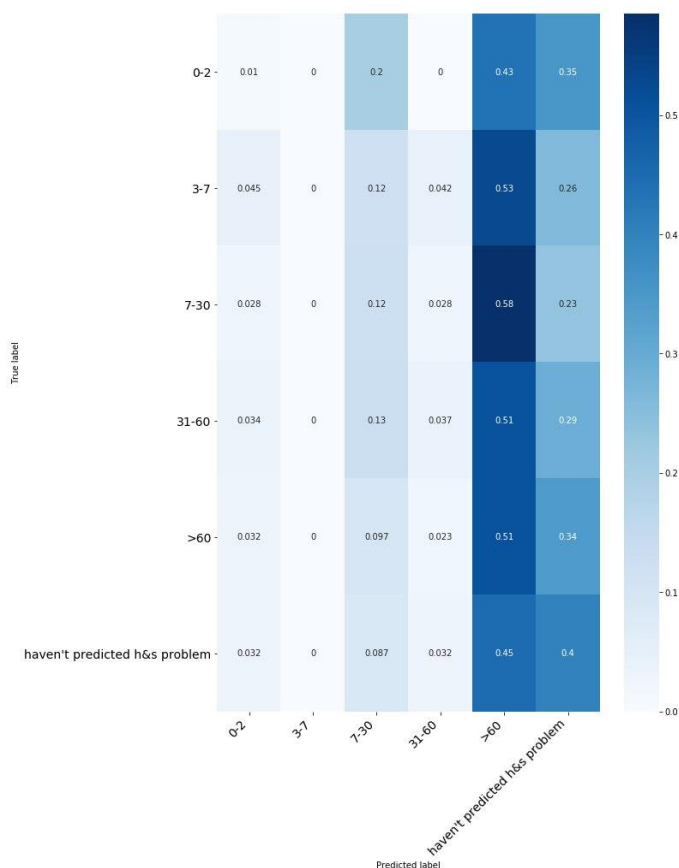


Рисунок 12 – Матрица несоответствия второго метода.

3.5. Выводы по данному решению

- Проведен анализ подхода и разработан алгоритм, способный решить стоящую перед нами задачу;
- Подобрана модульная архитектура, позволяющая разбить задачу на логические блоки;
- Создан модуль предварительной обработки текста;
- Создан и обучен модуль анализа текстовых данных;
- Собрана общая модель предсказаний;
- Модель предсказаний обучена и протестирована.

Созданная модель, классифицирует пациентов по тому, через сколько времени у них есть высокий риск возникновения сердечно сосудистых катастроф. Анализ проводится по последовательным данным, которые проходят через специально подготовленные структуры, находящие смысловые связи даже у удаленных друг от друга элементов. Результат не позволяет применять модель в данный момент, однако у сети есть потенциал улучшить показатели $recall$ и f меры до удовлетворительных показателей при более тонких настройках обучения.

4. Финансовый менеджмент, ресурсоэффективность, и ресурсосбережение

4.1. Потенциальные потребители результатов исследования

Разработанное программное обеспечение предугадывает потенциальную вероятность сердечно сосудистых катастроф у пациентов, на основании анализа данных из врачебных приемов. По данным ВОЗ около 22% людей находятся в группе риска, из-за чего проблема уменьшения тяжести течения кризиса крайне важно для здоровья всего населения. Разработка призвана на основе статистических данных и машинного обучения анализировать историю коммуникации человека с медицинским персоналом и предположить, с какой вероятностью и как скоро у человека возникнет ССК, позволяя заранее произвести профилактические мероприятия.

Программный комплекс разработан для выполнения специфической работы, поэтому потенциальный круг заинтересованных лиц ограничен. В первую очередь это медицинские учреждения, которые обслуживают большое количество пациентов. Так как программный комплекс способен предугадывать вероятность наступления сердечно сосудистых катастроф, в комплексе потенциально заинтересованы все государственные структуры, задачей которых является поддержание здоровья населения и снижения финансовых издержек на эти цели.

4.2. Технология QuaD

Произведем анализ по технологии QuaD (табл. 3).

Таблица 3 – Оценочная карта для сравнения конкурентных технических решений (разработок)

Критерии оценки	Вес критерия	Баллы	Макс. балл	Относительно е значение	Средневзвешенное значение
Показатели оценки качества разработки					
1. Качество сопровождающей документации	0,04	80	100	0,8	3,2
2. Восстанавливаемость	0,02	75	100	0,75	1,5
3. Безаварийность работы	0,05	95	100	0,95	4,75
4. Встраиваемость системы	0,05	90	100	0,9	4,5
5. Точность работы алгоритмов предсказания	0,2	80	100	0,8	16
6. Анализируемость	0,01	100	100	1	1
7. Изменяемость	0,06	95	100	0,95	5,7
8. Стабильность	0,02	65	100	0,65	1,3
9. Соответствие стандартам	0,06	85	100	0,85	5,1
10. Функциональная исправность	0,05	70	100	0,7	3,5
11. Интерфейс продукта	0,02	85	100	0,85	1,7
12. Качество кодовой базы	0,02	80	100	0,8	1,6
13. Наличие шифрования	0,02	95	100	0,95	1,9
14. Портативность программного продукта	0,03	75	100	0,75	2,25
15. Эффективность использования ресурсов	0,02	40	100	0,4	0,8
16. Эффективность по времени	0,01	65	100	0,65	0,65
17. Удобство изучения	0,1	75	100	0,75	7,5
18. Понятность	0,1	75	100	0,75	7,5
19. Удобство и простота использования	0,05	60	100	0,6	3
20. Совместимость	0,07	70	100	0,7	4,9
Итого	1	-	-	-	78,35

Оценка качества и перспективности по технологии QuaD определяется

по формуле:

$$P_{cp} = \sum B_i B_{i_i} \quad (1)$$

где $P_{ср}$ – средневзвешенное значение показателя качества и перспективности научной разработки;

$В_i$ – вес показателя (в долях единицы);

$Б_i$ – балл i -го показателя.

Тогда $P_{ср} = 78,35$.

Данное значение лежит в интервале от 60 до 79, следовательно, перспективность данной разработки выше среднего.

4.3.SWOT-анализ

Для изучения сильных и слабых сторон проекта требуется провести SWOT-анализ.

Таблица 4 – Матрица SWOT

	<p>Сильные стороны:</p> <p><i>С1. Принципиальная новизна продукта и отсутствие конкурентов.</i></p> <p><i>С2. Программный продукт не требует наложенных затрат на работу.</i></p> <p><i>С3. Интеллектуальная помощь в критической отрасли общественной жизни не потеряет актуальность длительное время.</i></p> <p><i>С4. Множественные пути развития продукта на основании проведенного исследования.</i></p>	<p>Слабые стороны:</p> <p><i>Сл1. Узкая специализация продукта.</i></p> <p><i>Сл2. Дальнейшие улучшения продукта и его обслуживание требует дорогостоящего сопровождения квалифицированным персоналом.</i></p> <p><i>Сл3. Регулирование области законодательством накладывает ограничения на спектр применяемых решений.</i></p>
<p>Возможности:</p> <p><i>В1. Доведение исследовательского прототипа до рабочей точности и внедрение продукта в работу.</i></p> <p><i>В2. Расширение структур заинтересованных в продукте.</i></p>	<p>Демонстрация заинтересованным лицам возможностей и удобства продукта, перспективы развития.</p>	<p>Увеличение возможностей продукта для расширения сферы применения.</p> <p>Нахождение возможностей продвижения проекта существующими структурами.</p>
<p>Угрозы:</p> <p><i>У1. Бюрократические ограничения разработки и её продвижения и внедрения.</i></p> <p><i>У2. Отсутствие ресурсов на адаптацию продукта к изменяющимся реалиям и его деградация.</i></p>	<p>Решение существующих проблем за счет современных технологий позволит улучшить экономическую эффективность пользователей, что позволит увеличить заинтересованность государства и преодолеть бюрократические барьеры.</p>	<p>При выдающихся результатах на пробных испытаниях, возможно, нахождения финансирования дальнейшего вектора развития.</p>

4.4. Планирование научно-исследовательских работ

4.4.1. Структура работ в рамках научного исследования

Рабочая группа состоит из людей, непосредственно имеющих отношение к разработке программного продукта – это разработчик и научный руководитель. Этапы разработки представлены в таблице 5.

Таблица 5 – Перечень этапов, работ и распределение исполнителей

№ работы	Наименование работы	Исполнители работы
1	Выбор научного руководителя бакалаврской работы	Мнацаканян Рафаэль Эдуардович
2	Составление и утверждение темы бакалаврской работы	Кудинов Антон Викторович, Мнацаканян Рафаэль Эдуардович
3	Составление календарного плана-графика выполнения бакалаврской работы	Кудинов Антон Викторович
4	Подбор и изучение литературы по теме бакалаврской работы	Мнацаканян Рафаэль Эдуардович
5	Анализ предметной области	Мнацаканян Рафаэль Эдуардович
6	Разбиение абстрактной задачи на спектр реальных задач	Мнацаканян Рафаэль Эдуардович
7	Проработка путей решения поставленных задач	Мнацаканян Рафаэль Эдуардович
8	Проведение исследования для понимания возможности решения основной задачи каждым путем	Мнацаканян Рафаэль Эдуардович
9	Подготовка решения задачи и анализ метрик	Мнацаканян Рафаэль Эдуардович
10	Согласование выполненной работы с научным руководителем	Кудинов Антон Викторович, Мнацаканян Рафаэль Эдуардович
11	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Мнацаканян Рафаэль Эдуардович
12	Подведение итогов, оформление работы	Мнацаканян Рафаэль Эдуардович

4.4.2. Определение трудоемкости выполнения работ

Трудовые затраты в большинстве случаев образуют основную часть стоимости разработки, поэтому важным моментом является определение трудоемкости работ каждого из участников научного исследования.

Трудовые затраты определяется по формуле:

$$t_{ож,i} = \frac{3t_{min,i} + 2t_{max,i}}{5}, \quad (2)$$

где $t_{ож,i}$ – ожидаемая трудоемкость выполнения i -ой работы, чел.-дн.;

$t_{min,i}$ – минимально возможная трудоемкость выполнения i -ой работы (оптимистическая оценка: в предположении наиболее благоприятного стечения обстоятельств), чел.-дн.;

$t_{max,i}$ – максимально возможная трудоемкость i -ой работы (пессимистическая оценка: в предположении наиболее неблагоприятного стечения обстоятельств), чел.-дн.

Помимо трудовых затрат, требуется рассчитать продолжительность работы в рабочих днях и продолжительность работы в календарных днях.

Формула для расчета продолжительности работы в рабочих днях:

$$T_{p,i} = \frac{t_{ож,i}}{ч_i}, \quad (3)$$

где $T_{p,i}$ – продолжительность одной работы, раб. дн.;

$t_{ож,i}$ – ожидаемая трудоемкость выполнения одной работы, чел.-дн.

$ч_i$ – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

Формула для расчета продолжительности работы в календарных днях:

$$T_{к,i} = T_{p,i} k_{кал}, \quad (4)$$

где $T_{к,i}$ – продолжительность выполнения i -ой работы в календарных днях;

$T_{p,i}$ – продолжительность выполнения i -ой работы в рабочих днях;

$k_{кал}$ – коэффициент календарности.

Коэффициент календарности рассчитывается по формуле:

$$T_k = \frac{T_{\text{кал}}}{T_{\text{кал}} - T_{\text{вд}} - T_{\text{пд}}}, \quad (5)$$

где $T_{\text{кал}}$ – календарные дни;

$T_{\text{вд}}$ – выходные дни;

$T_{\text{пд}}$ – праздничные дни.

Согласно производственному календарю на 2019 г:

$$T_{\text{кал}} = 365;$$

$$T_{\text{вд}} + T_{\text{пд}} = 66 \text{ дней.}$$

Тогда:

$$k_{\text{кал}} = \frac{365}{365 - 66} = 1,22.$$

Результаты расчетов представлены в сводной таблице временных показателей научного исследования в таблице 6.

На основе полученных данных строится календарный план график. Берется то исполнение, которое удовлетворяет требованиям заказчика. Продолжительность выполнения работы – с марта по июнь включительно

Таблица 6 – Временные показатели проведения научного исследования

Наименование работы	Исполнители работы	Трудоемкость работ, чел-дни			Длительность работ, дни	
		tmin	tmax	тож	Тр	Тк
Выбор научного руководителя бакалаврской работы	Мнацаканян Рафаэль Эдуардович	1	1	1	1	2
Составление и утверждение темы бакалаврской работы	Кудинов Антон Викторович	1	1	1	1	2
	Мнацаканян Рафаэль Эдуардович	2	5	4	4	5
Составление календарного плана-графика выполнения бакалаврской работы	Кудинов Антон Викторович	1	2	2	2	3

Продолжение таблицы 6

Подбор и изучение литературы по теме бакалаврской работы	Мнацаканян Рафаэль Эдуардович	20	60	36	36	44
Анализ предметной области	Мнацаканян Рафаэль Эдуардович	10	30	18	18	22
Разбиение абстрактной задачи на спектр реальных задач	Мнацаканян Рафаэль Эдуардович	3	5	4	4	5
Проработка путей решения поставленных задач	Мнацаканян Рафаэль Эдуардович	10	20	14	14	18
Проведение исследования для понимания возможности решения основной задачи каждым путем	Мнацаканян Рафаэль Эдуардович	20	50	32	32	40
Подготовка решения задачи и анализ метрик	Мнацаканян Рафаэль Эдуардович	12	20	16	16	20
Согласование выполненной работы с научным руководителем	Кудинов Антон Викторович	3	10	6	6	8
	Мнацаканян Рафаэль Эдуардович	3	10	6	6	8
Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Мнацаканян Рафаэль Эдуардович	5	10	7	7	9
Подведение итогов, оформление работы	Мнацаканян Рафаэль Эдуардович	5	10	7	7	9

Для большей наглядности временного ряда проведения научного исследования был построен график Ганта, рисунок 13.

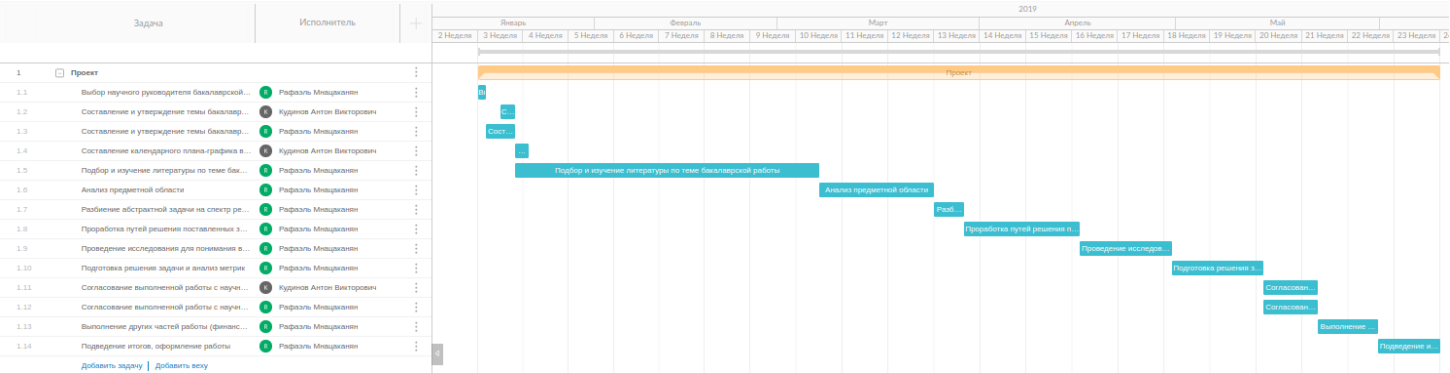


Рисунок13 – График Гантта

4.5. Бюджет научно-технического исследования

При планировании бюджета НТИ должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением. В процессе формирования бюджета НТИ используется следующая группировка затрат по статьям:

затраты на специальное оборудование для научных (экспериментальных) работ;

основная заработная плата исполнителей темы;

дополнительная заработная плата исполнителей темы;

отчисления во внебюджетные фонды (страховые отчисления);

Затраты на электроэнергию и прочие расходы.

4.5.1. Расчет затрат на специальное оборудование для научных работ

Для реализации проекта: создания документации, программного обеспечения и обучения математических моделей, необходимо два персональных компьютера с высокопроизводительными графическими ускорителями. Также стоит учитывать равномерное списание стоимости амортизируемого имущества в течении срока эксплуатации – линейный способ начисления амортизации.

С учетом срока службы ПК в 3 года (срок полезного использования для машин офисных код 330.28.23.23) и периодом разработки в 6 месяцев рассчитаем амортизационные отчисления.

Норма амортизации:

$$K = \frac{1}{n} * 100 \quad (6)$$

где К – норма амортизации в процентах к первоначальной (восстановительной) стоимости объекта амортизируемого имущества;

n – срок полезного использования данного объекта амортизируемого имущества, выраженный в годах.

Тогда,

$$K = \frac{1}{3} * 100\% = 33.33\%.$$

Сумма годовых амортизационных отчислений составит:

$$60000 * 0.33 = 20000(\text{руб.}).$$

За период разработки (5 мес.):

$$\frac{20000}{12} * 5 = 8333,33(\text{руб.}).$$

Так как было использовано две такие машины, общая величина амортизационных отчислений увеличивается в два раза.

$$8333,33 * 2 = 16666,67(\text{руб.}).$$

Таким образом, расчет затрат на приобретение оборудования представлен в таблице 7.

Таблица 7 – Специальное оборудование

№ п/п	Наименование оборудования	Количество единиц оборудования, шт.	Цена единицы оборудования, руб.	Общая стоимость оборудования, руб.
1.	Tensorflow framework	1	бесплатно	бесплатно
2.	Keras framework	1	бесплатно	бесплатно
3.	Libre Office	1	бесплатно	бесплатно
Амортизационные отчисления				16666,67
Итого				16666,67

4.5.2. Основная заработная плата исполнителей темы

В рамках реализации проекта руководитель и разработчик получают заработную плату.

Расчет заработной платы производится исходя из действительного годового фонда рабочего времени (таблица 8), оклада работника и длительности работ по трем исполнениям.

Таблица 8 – Баланс рабочего времени

Показатели рабочего времени	Руководитель	Разработчик
Календарное число дней	365	365
Количество нерабочих дней Выходные и праздничные дни	66	66
Потери рабочего времени Отпуск и невыходы по болезни	56	56
Действительный годовой фонд рабочего времени	243	243

Согласно таблице окладов ППС и НС, оклад доцента степени кандидата технических наук составляет 33664,00 руб., оклад разработчика – 21760,00 руб. без учета районного коэффициента.

Затраты на заработную плату:

$$З_{п} = З_{осн} + З_{доп} \quad (7)$$

$З_{осн}$ – основная заработная плата, руб.

$З_{доп}$ – дополнительная заработная плата, руб.

Заработная плата основная:

$$З_{осн} = З_{дн} * T_p * (1 + K_{пр} + K_d) * K_p \quad (8)$$

$З_{дн}$ – среднедневная заработная плата, руб.

$K_{пр}$ – премиальный коэффициент (0,3);

K_d – коэффициент доплат и надбавок (0,2);

K_p – районный коэффициент (для Томска 1,3);

T_p – продолжительность работ, выполняемых работником, раб. дни.

$$З_{дн} = \frac{З_{м*М}}{F_d} \quad (9)$$

Z_m – месячный оклад работника, руб.

M – количество месяцев работы без отпуска в течение года:

при отпуске в 48 раб. дней $M=10,4$ месяца, 6-дневная неделя;

F_d – действительный годовой фонд рабочего времени персонала, раб. дни.

Зная месячную заработную плату каждого участника проекта, можно рассчитать соответствующую среднедневную заработную плату. Количество месяцев работы без отпуска принимается равным 10,4 (считается отпуск длиной 48 рабочих дня при 6-дневной рабочей недели):

$$Z_{\text{дн}}^{\text{рук}} = \frac{33664,20 * 10,4}{243} = 1440,77(\text{руб.}),$$

$$Z_{\text{дн}}^{\text{разр}} = \frac{21760,00 * 10,4}{243} = 931,29(\text{руб.}).$$

Основная заработная плата работника – произведение среднедневной зарплаты на количество рабочих дней, затраченных на реализацию проекта, по трём исполнениям. Результаты расчетов приведены в таблице 9.

Таблица 9 – Расчёт основной заработной платы

Исполнитель	$Z_{\text{дн}}$, руб.	$k_{\text{пр}}$	$k_{\text{д}}$	$k_{\text{р}}$	$T_{\text{р}}$	$Z_{\text{осн}}$
Научный руководитель	1440,77	0,3	0,2	1,3	9	25285,51
Разработчик	931,29				145	263322,20

4.5.3. Дополнительная заработная плата исполнителей темы

С учётом основной заработной платы, можно посчитать дополнительную заработную плату в размере 12 % от основной.

$$З_{\text{доп}}^{\text{рук}} = \frac{25285,51 * 12}{100} = 3034,26(\text{руб.}),$$

$$З_{\text{доп}}^{\text{разр}} = \frac{263322,20 * 12}{100} = 31598,67(\text{руб.}).$$

На основе полученных результатов рассчитаем полную заработную плату.

$$З_{\text{полн}}^{\text{рук}} = 25285,51 + 3034,26 = 28319,78(\text{руб.}),$$

$$З_{\text{полн}}^{\text{разр}} = 263322,20 + 31598,67 = 294920,90(\text{руб.}).$$

4.5.4. Отчисления во внебюджетные фонды

Здесь отражаются обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования, пенсионного фонда и медицинского страхования от затрат на оплату труда.

Отчисления рассчитываются, как основная и дополнительная заработная плата сотрудника, умноженная на коэффициент отчислений. Для работников научно-образовательных учреждений данный коэффициент устанавливается равным 30%.

$$\text{ОТЧ}_{\text{п}}^{\text{рук}} = \frac{28319,78 * 30}{100} = 8495,93(\text{руб.}),$$

$$\text{ОТЧ}_{\text{п}}^{\text{разр}} = \frac{294920,90 * 30}{100} = 88476,28(\text{руб.}).$$

4.5.5. Накладные расходы

Во время производственного процесса непременно возникают так же затраты, не учтенные в предыдущих пунктах. Это затраты на обслуживание инфраструктуры, связи, аренды помещений и прочие, которые не влияют на процесс напрямую, но без которых он был бы затруднен или невозможен.

Величина накладных расходов – 16% от (затрат на специальное оборудование, затрат на основную заработную плату, затрат на дополнительную заработную плату и страховых взносов).

$$\text{Нак} = (16666,67 + 25285,51 + 263322,20 + 3034,26 + 31598,67 + 8495,93 + 88476,28) * 0,16$$

$$\text{Нак} = 69900,72(\text{руб.})$$

4.5.6. Формирование бюджета затрат научно-исследовательского проекта

Представим расчет бюджета на научно-исследовательскую работу в таблице 10.

Таблица 10 – Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб.	Удельный вес, %
1. Затраты на специальное оборудование для научных (экспериментальных) работ	16666,67	2,29%
2. Затраты по основной заработной плате исполнителей темы	25285,51+ 263322,20	56,95%
3. Затраты по дополнительной заработной плате исполнителей темы	3034,26 + 31598,67	6,83%
4. Страховые взносы	8495,93 + 88476,28	19,13%
5. Накладные расходы	69900,72	13,79%
6. Общий бюджет	506780,20	100%

4.6.Определение потенциального эффекта исследования

Исходя из расчетов и анализов, проведенных в данной части бакалаврской работы, выявлены сильные и слабые стороны проекта с экономической точки зрения. Фактическая уникальность разработки и рассчитанное значение перспективности научной разработки в 78,35 позволяет заключить, что разработка перспективна. Для успешного выхода на рынок необходимо обеспечить высокую точность предсказаний и низкий уровень ложных срабатываний. Лучшие результаты позволят снизить издержки потребителей, что позволит повысить экономическую рентабельность.

Вариант, разработки, принятый для реализации в итоге имеет бюджет 506780,20 рублей, основными статьями расходов являются заработная плата и страховые взносы, что предсказуемо для разработки ПО. Наибольшие материальные затраты связаны с обслуживанием двух ЭВМ с наличием производительных графических ускорителей.

5. Социальная ответственность

Целью выпускной квалификационной работы является создание программного обеспечения, которое будет предсказывать вызовы скорой помощи. Сервис представляет собой десктопное приложение, следовательно, разработка и эксплуатация данного сервиса предполагает работу в помещении оператором ЭВМ.

Данный раздел содержит описание экологической и производственной безопасности при разработке и использовании сервиса, также в нём рассмотрены вероятные ситуации опасного и чрезвычайного характеров, предложены способы их профилактики.

Понятие «социальная ответственность» включает в себя вопросы соблюдения прав персонала на труд, выполнения требований к обеспечению безопасности труда, охране окружающей среды и снижению вредных воздействий на неё.

Целью данного раздела является принятие проектных решений, исключающих несчастные случаи в производстве, а также снижение вредных воздействий на окружающую среду.

5.1. Правовые и организационные вопросы обеспечения безопасности

5.1.1. Специальные правовые нормы трудового законодательства

Нормы трудового законодательства регулируют отношения, связанные с использованием личного труда. Организация рабочего процесса обязательно должна осуществляться в соответствии с Трудовым кодексом Российской Федерации.

Согласно Трудовому кодексу Российской Федерации, а также инструкции по охране труда, организация труда должна происходить согласно следующим пунктам:

- рабочая смена должна составлять не более 8 часов;

- продолжительность непрерывной работы с компьютером без регламентированного перерыва не должна превышать 2-х часов (устанавливаются перерывы продолжительностью 15 минут);
- обеденный перерыв должен составлять не менее 30 минут и не более 1 часа.

Перед приёмом сотрудника на работу работодатель обязан провести инструктаж по технике безопасности, а также обеспечить периодическое проведение инструктажа в дальнейшем. Одной из обязанностей работодателя также является обеспечение сотрудников соответствующим нормам рабочим местом:

- рабочее место сотрудника должно быть организовано с учетом эргономических требований согласно ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования»;
- конструкция рабочей мебели должна предусматривать возможность индивидуальной регулировки (т.е. адаптироваться под рост и другие индивидуальные особенности сотрудника).

Работодатель обязан предпринимать меры, необходимые для профилактики производственного травматизма, профессиональных и других видов заболеваний работников.

5.1.2. Организационные мероприятия по компоновке рабочей зоны

Рабочее место сотрудника, использующего персональный компьютер, должно быть организовано в соответствии с требованиями, указанными в ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования» и СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы». Данные документы описывают нормы и требования к организации рабочего места и помещения, в котором оно располагается.

Общие требования к организации рабочего места сотрудника, работа которого связана с использованием персонального компьютера:

- экран монитора должен находиться от глаз пользователя на расстоянии 600 – 700 мм, но не ближе 500 мм с учетом размеров отображаемых на экране знаков и символов;
- деятельность с персональным компьютером должна чередоваться с перерывами, в которые пользователь не будет использовать персональный компьютер во избежание утомления пользователя;
- конструкция рабочего стула (кресла) должна обеспечивать поддержание рациональной рабочей позы при работе с компьютером. Рабочий стул (кресло) должен быть подъемно-поворотным, регулируемым по высоте и углам наклона сиденья и спинки.
- поверхность сиденья, спинки и других элементов стула (кресла) должна быть полумягкой, с нескользящим, слабо электризующимся и воздухопроницаемым покрытием, обеспечивающим легкую очистку от загрязнений.

5.2.Производственная безопасность

Разработка сервиса происходила в помещении с рабочим местом, оборудованным персональным компьютером с ЖК-дисплеем. В данном разделе рассмотрены вредные и опасные факторы, приведены рекомендации по их устранению.

Вредным является производственный фактор, воздействие которого на работающего в определенных условиях может привести к заболеванию, снижению работоспособности и (или) отрицательному влиянию на здоровье потомства. Опасным – фактор, воздействие которого на работающего в определенных условиях приводит к травме, острому отравлению или другому внезапному резкому ухудшению здоровья, или смерти.

В таблице 11 представим все опасные и вредные факторы, с которыми столкнутся пользователи и проектировщики системы и нормативная

документация, в которой описаны значения факторов, при которых рабочие условия безопасны и пригодны для описанного рабочего места.

Таблица 11 – Возможные опасные и вредные факторы

Наименование работ	Факторы (ГОСТ 12.0.003-2015)	Нормативные документы
Разработка и эксплуатация модуля системы с использованием ЭВМ	Вредные факторы	
	Отклонение показателей микроклимата рабочей зоны (температура и влажность воздуха)	СанПиН 2.2.2/2.4.1340-03, СанПиН 2.2.4.548-96
	Недостаточная освещенность рабочей зоны	СанПиН 2.2.2/2.4.1340-03
	Повышенный уровень электромагнитных излучений	СанПиН 2.2.2/2.4.1340-03, СанПиН 2.2.4.1191-03
	Повышенный уровень шума	СанПиН 2.2.4/2.1.8.562-96
	Опасные факторы	
	Опасность поражения электрическим током	ГОСТ 12.1.038-82 ГОСТ Р 22.0.07-95
	Пожарная безопасность	ФЗ от 22.07.2008 №123-ФЗ

5.2.1. Вредные факторы

5.2.1.1. Электромагнитное излучение

Длительное воздействие электромагнитных излучений, источником которых является ПК, может стать причиной снижения иммунитета, возникновения мигреней, ухудшения памяти, а также стать катализатором развития серьезных заболеваний.

Большое количество компьютерной техники в помещении, к примеру, в компьютерном классе, приводит к возникновению повышенного уровня электромагнитного излучения.

Допустимый уровень электромагнитного излучения устанавливается в СанПиН 2.2.2/2.4.1340-03. Защита от электромагнитного излучения может быть осуществлена установлением предельно допустимого уровня

напряженности, который, согласно СанПиН 2.2.2/2.4.1340-03, составляет не более 8 кА/м, при этом уровень магнитной индукции составляет 10 мТл.

В это же время, максимально допустимый уровень напряженности достигает 2,5 В/м, а плотность магнитного потока ограничена 25 нТл по тому же СанПиН, при условии работы высокочастотных излучателей.

5.2.1.2. Повышенный уровень шума

Источниками шума на рабочем месте могут являться системный блок персонального компьютера, устройства вентиляции, кондиционирования помещения, оргтехника и люди, находящиеся в помещении.

Ненормированные показатели шума на рабочем месте оказывают психологическое влияние на состояние сотрудника, вследствие чего падает уровень концентрации и сосредоточенности, понижается эффективность выполнения задач. Уровень шума также влияет на уровень стресса сотрудника и его утомляемость. Длительное воздействие шума может привести к снижению слуха.

Во избежание перечисленных выше последствий воздействия шума, требуется соблюдать установленные в СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы» требования. Согласно нормам, представленным в этом документе, уровень шума не должен превышать 50 дБА.

Для снижения уровня шума могут быть использованы следующие средства:

- устройства вентиляции и кондиционирования со сниженным уровнем шума;
- звукопоглощающий корпус ПК;
- вентиляторы охлаждения корпуса ПК со сниженным уровнем шума;
- Звукоизоляция помещения;

5.2.1.3. Недостаточная освещенность рабочей зоны

Освещение рабочего места имеет большое значение и оказывает влияние на работу сотрудника, а также на его физическое состояние. Значение имеет как естественное, так и искусственное освещение. Недостаток освещения ведет к ухудшению зрения работника

Согласно СанПиН 2.2.2/2.4.1340-03, освещенность на поверхности рабочего стола пользователя персонального компьютера должна быть 300 – 500 лк. Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк.

Помимо этого, существуют некоторые общие рекомендации и требования к организации освещения на рабочем месте, например:

- рабочие столы следует размещать таким образом, чтобы мониторы были ориентированы боковой стороной к световым проемам, а естественный свет падал преимущественно слева;
- искусственное освещение в помещениях для эксплуатации компьютера должно осуществляться системой общего равномерного освещения.

Соблюдение вышеуказанных мер позволит избежать негативного влияния на зрение работника.

5.2.1.4. Статические физические нагрузки

Статическое мышечное напряжение, возникающее при работе с персональным компьютером, возникает из-за сидячего положения пользователя. Концентрация внимания на действиях, происходящих на мониторе компьютера, приводят к перенапряжению мышц шейного отдела позвоночника, вследствие чего увеличивается риск возникновения сколиоза, остеохондроза и ухудшения мозгового кровообращения.

Возникновению подобных проблем способствует оснащение рабочего места нерегулируемой мебелью (отсутствует возможность регулировать высоту стула и рабочей области), а также отсутствием подлокотников и т.п.

Нормы организации рабочего пространства пользователя ПК регулируются в соответствии с СанПиН 2.2.2/2.4.1340-03.

Для того, чтобы минимизировать влияние статических нагрузок при работе с компьютером, требуется оснастить рабочее место правильно подобранной мебелью, соблюдать режимы работы и отдыха, а также выполнять ряд специальных физических упражнений.

5.2.1.5. Перенапряжение глаз

Наблюдение за действиями, происходящими на экране компьютера, приводит к повышению нагрузки на глаза пользователя. Ввод текста также является причиной перенапряжения глаз, так как пользователю приходится переносить свой взгляд с клавиатуры на монитор и обратно.

Длительная работа с ПК приводит к ежедневному утомлению глаз и к ухудшению зрения в длительной перспективе.

Минимизация влияния указанного фактора может быть произведена через чередование режимов работы и отдыха. При этом перерывы в работе должны производиться без использования ПК. Согласно СанПиН 2.2.2/2.4.1340-03, суммарное время перерывов должно составлять не менее 90 минут при 8-ми часовой рабочей смене.

Для профилактики перенапряжения может проводиться зрительная гимнастика, способствующая расслаблению и восстановлению глаз.

5.2.2. Опасные факторы

5.2.2.1. Опасность поражения электрическим током

Пользователь персонального компьютера взаимодействует непосредственно с электрооборудованием, поэтому всегда есть вероятность поражения электрическим током. В результате поражения током человек

может получить механические повреждения, возникающие из-за сокращения мышц под действием тока и ожоги.

Для того, чтобы избежать получения травм, пользователю персонального компьютера следует проверить свое рабочее место перед началом работы: убедиться, что розетки закреплены надежно и отсутствуют оголенные провода или видимые признаки неисправности оборудования. В случае обнаружения неисправностей, нельзя начинать работу с оборудованием или пытаться самостоятельно исправить неполадки. Неисправности должны быть устранены квалифицированным специалистом.

5.2.2.2. Опасность возникновения короткого замыкания

Короткое замыкание возникает при высоком уровне напряжения в сети и может спровоцировать пожар. Опасность представляет возможное поражение электрическим током.

Мерой предосторожности является использование стабилизатора напряжения или сетевого фильтра, которые позволяют защитить оборудование от перепадов напряжения. Необходимо также, чтобы электропроводка была скрытой.

5.2.2.3. Повышенный уровень статического электричества

Опасность повышенного уровня статического электричества состоит в том, что оно может вызывать головные боли, чрезмерную раздражительность и эмоциональность у работника, а также нарушение сна. Наибольшая опасность статического электричества состоит в возможности возникновения быстрого искрового разряда между частями оборудования. Такой заряд может привести к выходу оборудования из строя, электрическим травмам у работника или же к возникновению пожара.

Значение показателей уровня напряженности электростатических полей на рабочем месте регулируется ГОСТ 12.1.045–84 ССБТ «Электростатические поля. Допустимые уровни на рабочих местах и требования к проведению контроля». В этом документе предельно

допустимый уровень напряженности электростатических полей составляет 60 кВ/м в течение 1 часа.

Для минимизации количества статического электричества следует заземлять оборудование и коммуникации, а также повышать уровень влажности воздуха в помещениях.

5.3. Экологическая безопасность

Разрабатываемый проект представляет собой программу, поэтому негативное влияние на экологию окружающей среды при взаимодействии пользователя с продуктом связано с эксплуатацией персонального компьютера, а также с его утилизацией.

Негативное воздействие на окружающую среду может оказать неправильная утилизация батареи компьютера-ноутбука, которая содержит в себе тяжелые металлы, кислоты и щелочи, способные стать источником загрязнения литосферы или гидросферы. Для того, чтобы избежать возможного ущерба окружающей среде, при утилизации аккумуляторных батарей следует обращаться в специализированные учреждения, занимающиеся утилизацией и переработкой аккумуляторных батарей.

Другим опасным фактором воздействия на литосферу является неправильная утилизация люминесцентных ламп, которые используются для искусственного освещения на рабочих местах. Одно из опасных веществ, которое может загрязнить атмосферу, гидросферу и литосферу, а также стать причиной отравления человека и других живых существ – ртуть, содержащаяся в люминесцентных лампах в объеме от 10 до 70 мг. По истечении срока службы ламп (пять лет) их требуется сдавать на утилизацию в специализированные учреждения. Утилизация и транспортировка ламп должна производиться в соответствии с ГОСТ 6825-91.

5.4. Безопасность в чрезвычайных ситуациях

Работа с электроприборами, в частности, с персональным компьютером, связана с риском возникновения пожара. Пожар – это

неконтролируемый процесс горения, который происходит вне специального очага и наносит материальный ущерб, вред здоровью и жизни людей, интересам общества и государства.

Пожар может случиться из-за обрывания проводов, замыкания электропроводки оборудования, неисправности розеток или выключателей, при несоблюдении правил пожарной безопасности и т.д.

Существует основной ряд мер по предотвращению возникновения пожара:

- помещение требуется содержать в чистоте;
- должны быть свободны и ничем не загромождены проходы, лестничные клетки, коридоры и двери эвакуационных выходов;
- мебель и другие предметы не должны своим расположением препятствовать эвакуации в случае пожара;
- нельзя допускать использования неисправных электроприборов, розеток, рубильников, вилок и прочего оборудования;
- нельзя допускать курения в помещении и применения открытого огня;
- нельзя допускать использования электроприборов с поврежденной изоляцией.

В качестве мер предосторожности нужно ознакомиться с расположением огнетушителей и других средств пожаротушения, с планом эвакуации, а также убедиться в доступности представленных в плане эвакуационных выходов.

5.5.Выводы по разделу

В данном разделе были рассмотрены основные вредные и опасные факторы, которые могут возникнуть при работе с персональным компьютером. Также были предложены меры, которые позволят предотвратить негативное воздействие компьютера на человека. В качестве

вероятной чрезвычайной ситуации был представлен пожар. Меры противопожарной безопасности, а также требования к организации рабочего места и помещения были также рассмотрены в данной работе.

Помимо этого, раздел содержит основные нормы и требования как к рабочему месту, так и к порядку организации труда.

Рабочее место, использовавшееся при разработке программного сервиса, соответствует всем описанным нормам организации рабочего места.

Нормы организации труда регулируются, прежде всего, Трудовым кодексом Российской Федерации. Следует также отметить, что ни один внутренний регламент организации не должен противоречить Трудовому кодексу.

Исследование, проведённое при выполнении данного раздела, показало, что нет необходимости дополнительно модернизировать рабочее место пользователя или внедрять дополнительные меры безопасности при разработке и эксплуатации сервиса, достаточно выполнять приведённые выше рекомендации.

Заключение

В данной выпускной квалификационной работе были выполнены следующие задачи:

- изучена работа скорой медицинской помощи, работа, проводимая врачами в результате врачебных приемов и структурное наполнение их информационных систем;
- изучены алгоритмы машинного обучения для решения задач классификации;
- разработаны несколько перспективных подходов решения проблемы предотвращения повторных вызовов скорой помощи;
- созданы системы для проверки подходов и обучены модели;
- получены результаты метрик и сделаны выводы.

Таким образом задача создания системы предотвращения повторных вызовов скорой помощи выполнена частично.

Полученные модели позволяют с хорошей точностью предсказывать небольшое количество вызовов, что лучше полного отсутствия какой-либо системы предсказания, а при условии небольших расходов на работу, возможно её применение.

Проведены исследования, которые подтвердили эффективность некоторых подходов и найдены перспективные методы продолжения работы для улучшения точности до уровня возможности применения в реальных условиях.

В ходе выполнения данной работы был получен дополнительный опыт анализа данных, разработки алгоритмов машинного обучения и навыки работы со сторонним ПО.

Список используемых источников

1. Федеральный закон от 21.11.2011 N 323-ФЗ (ред. от 27.12.2018) "Об основах охраны здоровья граждан в Российской Федерации" (с изм. и доп., вступ. в силу с 31.01.2019) // consultant [Электронный ресурс]. http://www.consultant.ru/law/podborki/skoraya_medicinskaya_pomosch/
2. Приказ Министерства здравоохранения и социального развития РФ от 2 декабря 2009 г. N 942 «Об утверждении статистического инструментария станции (отделения), больницы скорой медицинской помощи» // consultant [Электронный ресурс]. http://www.consultant.ru/document/cons_doc_LAW_96009/
3. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.
4. Спицын В.Г. Применение искусственных нейронных сетей для обработки информации: Методические указания к лабораторным работам / Спицын В.Г., Цой Ю.Р. – Томск: Издательство ТПУ, 2007 – 32 с.
5. Применение нейронных сетей для задач классификации // basegroup [Электронный ресурс]. URL: <https://basegroup.ru/community/articles/classification>
6. Метод деревьев решений для задачи классификации // edu.kpfu [Электронный ресурс]. URL: https://edu.kpfu.ru/pluginfile.php/91556/mod_resource/content/3/Decision%20trees_1.pdf
7. Breiman L. Bagging predictors // Machine Learning, 1996, vol. 24, no. 2, pp. 123–140.
8. Leo Breiman. Random Forests // Machine Learning, October 2001, Volume 45, Issue 1, pp 5-32.

9. Caruana R., Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms // Department of Computer Science, Cornell University, Ithaca, NY 14853 USA.
10. Mark R. Segal. Machine Learning Benchmarks and Random Forest Regression // Division of Biostatistics, University of California, San Francisco, CA 94143-0560, April 14, 2003.
11. L. Mason , J. Baxter , P. Bartlett , M. Frea. Boosting Algorithms as Gradient Descent // 2000
12. Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. М.: ДМК Пресс, 2017. – 418 с.: ил.
13. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес. [Электронный ресурс]. URL: <https://habr.com/ru/company/ods/blog/324402/#3-ocenka-vazhnosti-priznakov>
14. Открытый курс машинного обучения. Тема 10. Градиентный бустинг. [Электронный ресурс]. URL: <https://habr.com/ru/company/ods/blog/327250/#4-itog-pro-teoriyu-gbm>
15. Метрики в задачах машинного обучения. [Электронный ресурс]. URL: <https://habr.com/ru/company/ods/blog/328372/>
16. СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы.
17. ГОСТ 12.4.011-89. Система стандартов безопасности труда. Средства защиты работающих. Общие требования и классификация. – М: ИПК издательство стандартов, 2001.
18. СанПиН 2.2.4.548-96. Гигиенические требования к микроклимату производственных помещений.
19. СанПиН 2.2.4/2.1.8.562-96. Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки.
20. ГОСТ 12.1.029-80. Средства и методы защиты от шума.
21. ГОСТ Р 55710-2013. Освещение рабочих мест внутри зданий.

22. ГОСТ 12.1.038-82 ССБТ. Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов.
23. СанПин 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы.
24. ГОСТ Р 53692-2009. Ресурсосбережение. Обращение с отходами. Этапы технологического цикла отходов.
25. ГОСТ 12.1.004-91. ССБТ Пожарная безопасность. Общие требования.
26. Davy Cielen, Arno D.B. Meysman, Mohamed Ali Introducing Data Sciencs – Manning Publications Co., 2016.
27. Конспекты курса МФТИ «Введение в машинное обучение», специализация – «Машинное обучение и анализ данных».